# An audio-visual approach to web video categorization

Bogdan Ionescu, Klaus Seyerlehner, Ionut Mironica, Constantin Vertan, Patrick Lambert

▶ **To cite this version:**

Bogdan Ionescu, Klaus Seyerlehner, Ionut Mironica, Constantin Vertan, Patrick Lambert. An audio-visual approach to web video categorization. Multimedia Tools and Applications, 2014, 70 (2), pp. 1007-1032. 10.1007/s11042-012-1097-x . hal-00732716

**HAL Id: hal-00732716**
**https://hal.science/hal-00732716**

Submitted on 17 Sep 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multimedia Tools and Applications

# An Audio-Visual Perspective on Automatic Web Media Categorization

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | |
| **Full Title:** | An Audio-Visual Perspective on Automatic Web Media Categorization |
| **Article Type:** | Multimedia on the Web |
| **Keywords:** | audio block-based descriptors; color perception; action assessment; video relevance feedback; video genre classification |
| **Corresponding Author:** | Bogdan Emanuel Ionescu, Ph.D.<br>University Politehnica of Bucharest<br>Bucharest, ROMANIA |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author's Institution:** | University Politehnica of Bucharest |
| **Corresponding Author's Secondary Institution:** | |
| **First Author:** | Bogdan Emanuel Ionescu, Ph.D. |
| **First Author Secondary Information:** | |
| **All Authors:** | Bogdan Emanuel Ionescu, Ph.D. |
| | Klaus Seyerlehner, Ph.D. |
| | Ionut Mironica, M.S. |
| | Constantin Vertan, Ph.D. |
| | Patrick Lambert, Ph.D. |
| **All Authors Secondary Information:** | |
| **Abstract:** | In this paper we address the issue of automatic video genre categorization of web media using an audio-visual approach. To this end, we propose content descriptors which exploit audio, temporal structure and color information. The potential of our descriptors is experimentally validated both from the perspective of a classification system and as an information retrieval approach. Validation is carried out on a real scenario, namely on more than 288 hours of video footage and 26 video genres specific to blip.tv media platform. Additionally, to reduce semantic gap, we propose a new relevance feedback technique which is based on hierarchical clustering. Experimental tests prove that retrieval performance can be significantly increased in this case, becoming comparable to the one obtained with high level semantic textual descriptors. |

*Powered by Editorial Manager® and Preprint Manager® from Aries Systems Corporation*

Manuscript
Click here to download Manuscript: MTAP_specialissue.tex
Click here to view linked References

**Noname manuscript No.**
(will be inserted by the editor)

# An Audio-Visual Perspective on Automatic Web Media Categorization

**Bogdan Ionescu · Klaus Seyerlehner · Ionuţ Mironică · Constantin Vertan · Patrick Lambert**

Received: date / Accepted: date

**Abstract** In this paper we address the issue of automatic video genre categorization of web media using an audio-visual approach. To this end, we propose content descriptors which exploit audio, temporal structure and color information. The potential of our descriptors is experimentally validated both from the perspective of a classification system and as an information retrieval approach. Validation is carried out on a real scenario, namely on more than 288 hours of video footage and 26 video genres specific to blip.tv media platform. Additionally, to reduce semantic gap, we propose a new relevance feedback technique which is based on hierarchical clustering. Experimental tests prove that retrieval performance can be significantly increased in this case, becoming comparable to the one obtained with high level semantic textual descriptors.

**Keywords** audio block-based descriptors · color perception · action assessment · video relevance feedback · video genre classification

B. Ionescu
LAPI, University "Politehnica" of Bucharest, 061071, Romania,
LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944, France,
E-mail: bionescu@alpha.imag.pub.ro

K. Seyerlehner
DCP, Johannes Kepler University, A-4040 Austria,
E-mail: klaus.seyerlehner@jku.at

I. Mironică
LAPI, University "Politehnica" of Bucharest, 061071, Romania,
E-mail: imironica@alpha.imag.pub.ro

C. Vertan
LAPI, University "Politehnica" of Bucharest, 061071, Romania,
E-mail: constantin.vertan@upb.ro

P. Lambert
LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944 France,
E-mail: patrick.lambert@univ-savoie.fr

## 1 Introduction

The automatic labeling of video footage according to genre is a common requirement when dealing with indexing of large and heterogenous collection of video materials. This task may be addressed, either *globally*, or *locally*. Global classification approaches aim to categorize videos into one of several main genres, e.g. cartoons, music, news, sports, documentaries; or even more fine-grained into sub-genres, e.g. identifying specific types of sports (e.g. football, hockey) or movies (e.g. drama, thriller). Local classification approaches, in contrast, label video segments instead of whole videos according to specific human-like concepts, e.g. outdoor vs. indoor scenes, action segments, violence scenes (see TRECVid campaign [3]).

In this paper we focus on the global classification task and video genre classification is consequently interpreted as a typical machine learning problem. In the literature so far, many sources of information have been exploited with respect to this task [1]. One of the most common approaches is the use of *text-based* information. Basically, all existing web media platform search engines (e.g. YouTube, blip.tv) rely on text-based retrieval as it accounts for a higher semantic level of description than other informative sources. Text is obtained either from scene text (e.g. graphic text, sub-titles), from the transcripts of dialogues obtained with speech recognition techniques or from other external sources, e.g. synopsis, user tags, metadata. Common genre classification approaches include the use of the classic Bag-of-Words model [4] or of the Term Frequency-Inverse Document Frequency (TF-IDF) approach [2].

Less accurate than text is the use of audio-visual information. *Audio-based* information may be derived from, both, time and frequency domains. Usual *time-domain* approaches include the use of Root Mean Square of signal energy (RMS) [23], sub-band information [5], Zero-Crossing Rate (ZCR) [25] or silence ratio; while *frequency-domain* features include energy distribution, frequency centroid [25], bandwidth, pitch [6] or Mel-Frequency Cepstral Coefficients (MFCC) [24].

The most popular type of audio-visual content descriptors are however the *visual descriptors*. They exploit, both static and dynamic visual information in the *spatial domain*, e.g. using color, temporal structure, objects, feature points, motion; or in the *compressed domain*, i.e. using MPEG coefficients [1]. *Color information* is generally derived at image level and is quantified via color histograms or other low-level parameters, e.g. predominant color, color entropy, variance (various color spaces are used, like RGB - Red Green Blue, HSV - Hue Saturation Value or YCbCr - Luminance, Chrominance) [7] [8]. *Temporal structure based information* exploits temporal segmentation of video sequences. A video sequence is composed of several video shots which are connected by video transitions, i.e. sharp transitions - cuts and gradual transitions - fades, dissolves, [26]. Existing approaches exploit basically their occurrence frequency in the movie. Although some approaches use directly this information [9] (e.g. rhythm, average shot length), commonly visual activity related features are derived by defining action content (e.g. a high frequency of

shot changes may be in general correlated to action) [19]. *Object-based features* used with genre classification are in general limited to the characterization of the occurrence of face and text regions in the frames [9] [19]. Other related approaches exploit the presence of feature points like the use of the well known SIFT descriptors [13]. *Motion-based information* is derived either from motion detection techniques (i.e. foreground detection) or from motion estimation (i.e. prediction of pixel displacement vectors from one frame to another). Common features describe motion density, camera movement (global movement) or object trajectory [11]. Finally, less common are features directly computed in the *compressed video domain*, e.g. using using DCT coefficients (Discrete Cosine Transform) or embedded motion vectors from MPEG flow [12]. Their main advantage is in the immediate availability with the video sequence.

All sources of information provide advantages and disadvantages, but however some prove to be more convenient than others depending on the classification scenario. *Text-based* information, due to its high informational redundancy and reduced availability with visual information, can be less relevant when addressing a reduced number of genres, e.g. TV media genres. Also, it can be subject to high error rates if retrieved with speech transcription techniques - however it holds the supremacy on addressing web genre categorization; *object-based* information, whether computational expensive it tends to be semi-automatic (requires human confirmation); *motion information* tends to be present in high quantities during the entire sequence (object/camera) and itself is not sufficient to distinguish between some genres, e.g. movies, sports, music. *Audio-based* information provides good discriminative power for most of the main TV media specific genres and require fewer computational resources to obtain and process; *color information* is simple to implement and inexpensive to process while powerful in distinguish for cinematic principles; *temporal-based* information is a popular choice and proves to be powerful as long as employing efficient video transition detection algorithms (e.g. adapting to web specific low-quality video contents [20]).

As far as *data categorization* techniques are concerned, some of the most popular choices with video genre classification are: Support Vector Machines (SVM), Bayesian-based, Neural Networks, Gaussian Mixtures and Hidden Markov Models (HMM) [1]. The remainder of the paper is organized as follows: Section 2 discuses several relevant genre classification approaches and situates our work accordingly. Section 3 presents the proposed video descriptors which are extracted from audio, temporal and color information. Section 4 discusses the improvement of the classification performance with relevance feedback and proposes an approach inspired by hierarchical clustering. Experimental results are presented in Section 5, while Section 6 presents the conclusions.

## 2 Related work

In this paper we focus on the audio-visual information as we aim to demonstrate its potential to web media genre categorization. Although, some sources

of information provide better results than others when applied to video genre classification [1], the most reliable approaches - which also target the wider range of genres - are however *multi-modal*, i.e. multi-source. In the following we shall highlight the performance of several approaches, from single-modal (which are limited to cope with a reduced number of genres) to multi-modal (able to perform more complex classifications) we consider relevant for the present work.

A simple, single modal approach is the one proposed in [14]. It addresses genre classification using only video dynamics. Motion information is extracted at two levels: background camera motion and foreground or object motion. A single feature vector is constituted in the DCT transformed space. This is to assure low-pass filtering, orthogonality and a reduced feature dimension. A Gaussian Mixture Model (GMM) based classifier is then used to identify 3 common genres: sports, cartoons and news. Despite the limited content information used, applied to a reduced number of genres achieves detection errors below 6%.

A much more complex approach which uses spatio-temporal information is proposed in [19]. At temporal level, video contents is described using average shot length, cut percentage, average color difference and camera motion (4 cases are detected: still, pan, zoom, and other movements). Spatial features include face frames ratio, average brightness and color entropy. Genre classification is addressed at different levels, according to a hierarchical ontology of video genres. Several classification schemes (Decision Trees and several SVM approaches) are used to classify video footage into main genres: movie, commercial, news, music and sports; and further into sub-genres, movies into action, comedy, horror and cartoon, and sports into baseball, football, volleyball, tennis, basketball and soccer. The highest precision for video footage categorization is around 88.6%, while for sub-genres, sports categorization achieve 97% and movies up to 81.3%.

A truly multi-modal approach, which combines several categories of content descriptors, is proposed in [29]. Features are extracted from four informative sources, which include visual-perceptual information (color, texture and motion), structural information (shot length, shot distribution, shot rhythm, shot clusters duration and saturation), cognitive information (face properties, such as number, positions and dimensions) and aural information (transcribed text, sound characteristics). These features are used for training a parallel Neural Network system and achieve an accuracy rate up to 95% in distinguish between seven video genres and sub-genres, namely: football (sub-genre of sports), cartoons, music, weather forecast (sub-genre of news), newscast, talk shows and commercials.

In this paper we propose three categories of content descriptors, which exploit both audio and visual modalities. Although these sources of information have already been exploited, one of the novelties of our approach is the way we compute the descriptors. The proposed *audio features* are block-level based, which compared to classic approaches, e.g. MFCC [12], have the advantage of capturing local temporal information by analyzing sequences of

consecutive frames in a time-frequency representation. *Visual information* is described with temporal information and color properties. Temporal descriptors are first derived using a classic confirmed approach, i.e. analyzing the frequency of shot changes [19] [29]. Then, we introduce a novel way to assess action content also considering the human perception. With respect to color information we aim to capture aspects of color perception. In contrast to typically low-level color descriptors, e.g. use of predominant color, color variance, color entropy or frame based histograms [19], we project histogram features onto a typical human color naming system and determine descriptors like the amount of light colors, cold colors, saturated colors, color contrasts or elementary hue distribution. This accounts for a higher semantic level of description. The proposed descriptors have been preliminary validated on the classification of seven common TV genres (animated movies, commercials, documentaries, movies, music videos, news broadcast and sports) leading to average precision and recall ratios within $[87\% - 100\%]$ and $[77\% - 100\%]$, respectively [16].

In this paper we extend and adapt the research to the categorization of web media genres. Several experimental tests conducted on a real scenario, namely using up to 26 genres provided with blip.tv media platform and more than 288 hours of video footage demonstrate the potential of these audio-visual descriptors for this task. Tests are conducted both from the perspective of a classification system and as an information retrieval approach.

If we approach this categorization problem from the perspective of a video retrieval system, there are two ways of improving its performance and thus to reduce semantic gap. One is to account for content descriptors as much as semantic as possible, as it is the case of textual information; or to compensate for this and take advantage directly of the user's expertise (i.e. "the consumer of the product") and adapt the system's response to his preferences. To this end, we propose a new relevance feedback technique which is based on hierarchical clustering. It allows to boost up the retrieval performance obtained with audio-visual descriptors close to the one obtained with high level semantic textual descriptors. All these aspects are presented in the sequel.

## 3 Content description

As previously presented, we approach video genre categorization with both, audio and visual information. From the existing modalities we exploit the *audio soundtrack*, *temporal structure* and *color distribution*.

Our selection is motivated by the specificity of these informative sources with respect to video genre. For instance, most of the common video genres have very specific audio signatures, e.g. music clips contain music, in news there are a lot of monologues/dialogues, documentaries have a mixture of natural sounds, speech and ambience music, in sports there is the specific crowd noise, an so on. This is to be found also at the visual level. Temporal structure and color information highlight specific genre contents, e.g. commercials and music clips tend to have a high visual tempo, music clips and movies tend

to have darker colors (mainly due to the use of special effects), commercials use a lot of gradual transitions, documentaries have a reduced action content, animated movies have specific color palettes and color contrasts, sports usually have a predominant hue (e.g. green for soccer, white for ice hockey), in news broadcasting there is the presence of the anchorman (high frequency of faces). In the following we shall present each category of descriptors and emphasize their advantages.

### 3.1 Audio descriptors

To address the specificity of video genres we propose audio descriptors which are related to rhythm, timbre, onset strength, noisiness and vocal aspects [18]. The proposed set of audio descriptors, called block-level audio features, have the key advantage of capturing also local temporal information from the audio track. Instead of using single frames only, temporal integration is carried out by analyzing sequences of consecutive frames (i.e. *blocks*) in a time-frequency representation. Blocks are of variable length and can be overlapping (e.g. by a maximum of 50% of their frames).

After converting the video soundtrack into a $22kHz$ mono signal, we compute short-time Fourier transform and perform a mapping of the frequency axis according to the logarithmic cent-scale. This is to account for the logarithmic human frequency perception. The resulting time-frequency representation consists of 97 logarithmically spaced frequency bands. Then, the following complex audio features are derived:

**- spectral pattern**: characterize the soundtrack's timbre via modeling those frequency components that are simultaneously active. Dynamic aspect of the signal are kept by sorting each frequency band of the block along the time axis. The block width varies depending on the extracted patterns, which allows to capture temporal information over different time spans.

**- delta spectral pattern**: captures the strength of onsets. To emphasize onsets, first we compute the difference between the original spectrum and a copy of the original spectrum delayed by 3 frames. Then, each frequency band is sorted along the time axis in a similar way as in the case of spectral pattern.

**- variance delta spectral pattern**: is basically an extension of the delta spectral pattern and captures the variation of the onset strength over time.

**- logarithmic fluctuation pattern**: captures the rhythmic aspects of the audio signal. In order to extract the amplitude modulations out of the temporal envelope in each band, periodicities are detected by computing the Fast Fourier Transform (FFT) along each frequency band of a block. The periodicity dimension is then reduced from 256 to 37 logarithmically space periodicity bins.

**- spectral contrast pattern**: roughly estimates the *"tone-ness"* of an audio track. For each frame, within a block, the difference between spectral peaks

and valleys in 20 sub-bands is computed and the resulting spectral contrast values are sorted along the time axis in each frequency band.

**- correlation pattern**. To capture the temporal relation of loudness changes over different frequency bands, we use the correlation coefficient among all possible pairs of frequency bands within a block. The resulting correlation matrix forms the so-called correlation pattern. The correlation coefficients are computed for a reduced frequency resolution of number of 52 bands.

These audio features in combination with a Support Vector Machine (SVM) classifier constitute a highly efficient automatic music classification system. During the last run of the Music Information Retrieval Evaluation eXchange, this approach ranked first with respect to the task of automatic music genre classification [18]. However, the proposed approach has not yet been applied to video genre classification. Existing approaches are limited to use standard audio features, e.g. a common approach is to use Mel-Frequency Cepstral Coefficients (MFCC) [24] or to use time domain features, e.g. Root Mean Square of signal energy (RMS) [23] or Zero-Crossing Rate (ZCR) [25].

### 3.2 Temporal structure descriptors

Temporal descriptors are derived using a classic confirmed approach, i.e. analyzing the frequency of shot changes [19]. Compared to existing approaches, we tune the assessment of action level based on human perception.

One of the main success factors of temporal descriptions is an accurate preceding temporal segmentation [26]. To this end we detect both cuts and also gradual transitions. Cuts are detected using an adaptation of the histogram-based approach proposed in [20], while fades and dissolves are detected using a pixel-level statistical approach [21] and the analysis of fading-in and fading-out pixels [22], respectively. Then, we compute the following descriptors:

**- rhythm**: to capture the movie's visual changing tempo, first we compute the relative number of shot changes occurring within a time interval of $T = 5s$, denoted $\zeta_T$. Then, the rhythm is defined as the movie average shot change ratio, namely $E\{\zeta_T\}$.

**- action**: we aim at highlighting two opposite situations: video segments with a high action content (denoted hot action, e.g. fast changes, fast motion, visual effects) with $\zeta_T > 3.1$, and video segments with low action content (i.e. containing mainly static scenes) with $\zeta_T < 0.6$. These thresholds were determined experimentally using user inputs. Several persons were asked to manually label video segments from a data set of ten relevant videos into the previous two categories. Using this information as ground truth, we determined average $\zeta_T$ intervals for each type of action content and following, the threshold limits.

Further, we quantify the action content with two parameters, hot-action ratio ($HA$) and low-action ratio ($LA$), thus:

$$HA = \frac{T_{HA}}{T_{total}}, \quad LA = \frac{T_{LA}}{T_{total}} \tag{1}$$

where $T_{HA}$ and $T_{LA}$ represent the total length of hot and low action segments, respectively, and $T_{total}$ is the movie total length.

**- gradual transitions ratio**: high amounts of gradual transitions are in general related to a specific video contents, therefore we compute:

$$GT = \frac{T_{dissolves} + T_{fade-in} + T_{fade-out}}{T_{total}} \tag{2}$$

where $T_X$ represents the total duration of all gradual transitions of type $X$. This provides information about editing techniques which are specific to certain genres, like movies or commercial clips.

### 3.3 Color descriptors

Color information is an important source to describe visual contents. Most of the existing color-based genre classification approaches are limited to use intensity-based parameters or generic low-level color features, e.g. average color differences, average brightness, average color entropy [19], variance of pixel intensity, standard deviation of gray level histograms, percentage of pixels having saturation above a certain threshold [27], lighting key [28], object color and texture.

We propose a more elaborated strategy which addresses the perception of the color content [31]. A simple and efficient way to accomplish this is with the help of color names; associating names with colors allows everyone to create a mental image of a given color or color mixture. We project colors on to a color naming system and colors properties are described using: statistics of color distribution, elementary hue distribution, color visual properties (e.g. amount of light colors, warm colors, saturated colors) and relationship of color (adjacency and complementarity). Prior to parameter extraction, we use an error diffusion scheme to project colors into a more manageable color palette, i.e. the non-dithering 216 color Webmaster palette (it provides an efficient color naming system). Color information is then represented with the following descriptors:

**- global weighted color histogram** is computed as the weighted sum of each shot color histogram, thus:

$$h_{GW}(c) = \sum_{i=0}^{M} \left[ \frac{1}{N_i} \sum_{j=0}^{N_i} h^j_{shot_i}(c) \right] \cdot \frac{T_{shot_i}}{T_{total}} \tag{3}$$

where $M$ is the total number of video shots, $N_i$ is the total number of the retained frames for the shot $i$ (we use temporal sub-sampling), $h^j_{shot_i}$ is the color histogram of the frame $j$ from the shot $i$, $c$ is a color index from the Webmaster palette (we use color reduction) and $T_{shot_i}$ is the length of the shot $i$. The longer the shot, the more important its contribution to the movie's global histogram.

**- elementary color histogram**: the next feature is the distribution of elementary hues in the sequence, thus:

$$h_E(c_e) = \sum_{c=0}^{215} h_{GW}(c)|_{Name(c_e) \subset Name(c)} \quad (4)$$

where $c_e$ is an elementary color from the Webmaster color dictionary (colors are named according to color hue, saturation and intensity) and $Name()$ returns a color's name from the palette dictionary.

**- color properties**: with this feature set we aim to describe color properties. We define several color ratios. For instance, light color ratio, $P_{light}$, reflects the amount of bright colors in the movie, thus:

$$P_{light} = \sum h_{GW}(c)|_{W_{light} \subset Name(c)} \quad (5)$$

where $c$ is a color with the property that its name contains one of the words defining brightness, i.e. $W_{light} \in$ {"light", "pale", "white"}. Using the same reasoning and keywords specific to each property, we define dark color ratio ($P_{dark}$), hard saturated color ratio ($P_{hard}$), weak saturated color ratio ($P_{weak}$), warm color ratio ($P_{warm}$) and cold color ratio ($P_{cold}$). Additionally, we capture movie color wealth with two parameters: color variation, $P_{var}$, which accounts for the amount of significant different colors and color diversity, $P_{div}$, defined as the amount of significant different color hues.

**- color relationship**: finally, we compute $P_{adj}$, the amount of similar perceptual colors in the movie and $P_{compl}$, the amount of opposite perceptual color pairs.

## 4 Relevance feedback

To improve the performance of the video genre categorization we investigate the potential of user Relevance Feedback techniques (RF). A general RF scenario can be formulated thus: for a certain retrieval query, user first marks some returned documents (documents are considered in a certain browsing window) as relevant or non-relevant; then the system computes a better representation of the information need based on this information and retrieval is further refined. Relevance feedback can go through one or more iterations of this sort [33]. This basically improves the system response based on query related ground-truth.

One of the earliest and most successful relevance feedback algorithms is the Rocchio algorithm [34]. It updates the query features by adjusting the position of the original query in the feature space according to the positive and negative examples and their associated importance factors. Another example is the Feature Relevance Estimation (RFE) approach [36] which assumes, for a given query, that according to the users subjective judgment, some specific features may be more important than others. Every feature will be given an

importance weight such that features with bigger variance have low importance than elements with low variations. More lately, Support Vector Machines found their application with relevance feedback approaches [37]. The problem can be formulated either as a two class classification of the negative and the positive samples; or as a one class classification problem, i.e. separate positive samples by negative samples.

To this end, we proposed a new RF method inspired from Hierarchical Clustering techniques (HC) [35]. Hierarchical clustering partition the observations into clusters. First, every document is assigned to a new cluster. During each iteration, hierarchical agglomerative clustering (HAC) successively searches for the most similar clusters in the current partition (based on computing measures like average distance, minimal variance, Ward's distance); these clusters are then merged resulting in the decrease of the total number of clusters by one. By repeating the process, HAC will produce a dendrogram of the observations, which may be informative for data display and discovery of data relationship. This clustering mechanism can be successfully exploited to the RF problem, and especially to our video genre retrieval issue where observations are the video sequences and the queries are related to video genre.

The proposed hierarchical clustering relevance feedback (HCRF) uses the general assumption that the video content descriptors provide enough representative power such that within the first window of retrieved video sequences are at least some relevant videos for the query that can be used as positive feedback (the retrieval results are divided into non-overlapping windows and RF is applied successively to each of them). This can be assured by considering the right size of the window. Also, in most cases, there is at least one non-relevant video that can be used as negative feedback. Instead of modifying the query or the similarity metric, as most RF algorithms do [33], we propose to simply cluster the remaining retrieved videos with respect to the genre label. At each feedback iteration, the retrieved videos that are in the following browsing window will be clustered with the following algorithm.

The first step consists on initializing the clusters using the video sequences in the current relevance feedback window. At the beginning, each cluster consists of a single video. Further, we perform the hierarchical aggregative clustering based on cluster centroid distance and by merging the most similar clusters within each relevance category (relevant and non-relevant). The clustering process stops when the number of clusters in the current iteration becomes relevant for the video genre categories within the browsing window (a heuristic choice is to set the minimal number of clusters equal to a quarter of the number of sequences). After finishing the training phase, we begin to classify the next videos as relevant or non-relevant with respect to the existing cluster repartition and by using the same reasoning as presented before. The entire process can be re-iterated if needed (e.g. retrieval performance is still low) by acquiring new relevance feedback information from the user.

The method's algorithm is presented with Algorithm 1 where the following notations were adopted: $N_{RV}$ is the number of sequences from the browsing window, $N_{clusters}$ represent the number of clusters, $sim[i][j]$ accounts for

**Algorithm 1** Hierarchical Clustering Relevance Feedback.

$N_{clusters} \leftarrow N_{RV}$; $clusters \leftarrow \{C_1, C_2, ..., C_{N_{clusters}}\}$;
**for** $i = 1 \rightarrow N_{clusters}$ **do**
  **for** $j = i \rightarrow N_{clusters}$ **do**
    *determine* $sim[i][j]$;
    $sim[j][i] \leftarrow sim[i][j]$;
  **end for**
**end for**
**while** $(N_{clusters} \geq \tau)$ **do**
  *determine* $argmin_{i,j}(sim[i][j])$;
  $N_{clusters} \leftarrow N_{clusters} - 1$;
  $C_{min} = C_{min_i} \cup C_{min_j}$;
  **for** $i = 1 \rightarrow N_{clusters}$ **do**
    *compute* $sim[i][min]$;
  **end for**
**end while**
$TP \leftarrow 0$; $current\_video \leftarrow N_{RD} + 1$;
**while** $((TP \leq \tau_1) \parallel (current\_video < \tau_2))$ **do**
  **for** $i = 1 \rightarrow N_{clusters}$ **do**
    *compute* $sim[i][current\_video]$;
  **end for**
  **if** ($current\_video$ is relevant) **then**
    $TP \leftarrow TP + 1$;
  **end if**
  $current\_video \leftarrow current\_video + 1$;
**end while**

the measure of similarity between clusters $C_i$ and $C_j$ (i.e. centroid distance), $\tau$, $\tau_1$ and $\tau_2$ are several thresholds which were empirically determined and *current_video* is the index of the current analyzed video.

## 5 Experimental results

The validation of the proposed content descriptors is carried out in the context of the MediaEval 2011 Benchmarking Initiative for Multimedia Evaluation, the Video Genre Tagging Task [2]. It addresses the automatic categorization of web media genres used with the blip.tv platform (see `http://blip.tv/`).

For testing we use a data set consisting of 2375 sequences (up to 288 hours of video footage) labeled according to 26 video genre categories, namely (the number in the brackets is the number of sequences): "*art*" (66), "*autos and vehicles*" (36), "*business*" (41), "*citizen journalism*" (92), "*comedy*" (35), "*conferences and other events*" (42), "*documentary*" (25), "*educational*" (111), "*food and drink*" (63), "*gaming*" (41), "*health*" (60), "*literature*" (83), "*movies and television*" (77), "*music and entertainment*" (54), "*personal of auto-biographical*" (13), "*politics*" (597), "*religion*" (117), "*school and education*" (11), "*sports*" (117), "*technology*" (194), "*environment*" (33), "*mainstream media*" (47), "*travel*" (62), "*videoblogging*" (70) and "*web development and sites*" (40) and "*default category*" (248, accounts for movies which cannot

be assigned to neither one of the previous categories). For more details on genre categories see [2].

The main challenge of this task is in the high number of genres, e.g. for our scenario up to 26. Also, each genre category has a high variety of video materials which may interfere with any training step. Last but not least, video contents available with media platforms is rather specific to video reports than to classic TV genres. Video materials are usually assembled in a news broadcasting style, i.e. genre specific contents is inserted periodically during a dialogue or interview scene. Figure 1 illustrates these aspects.

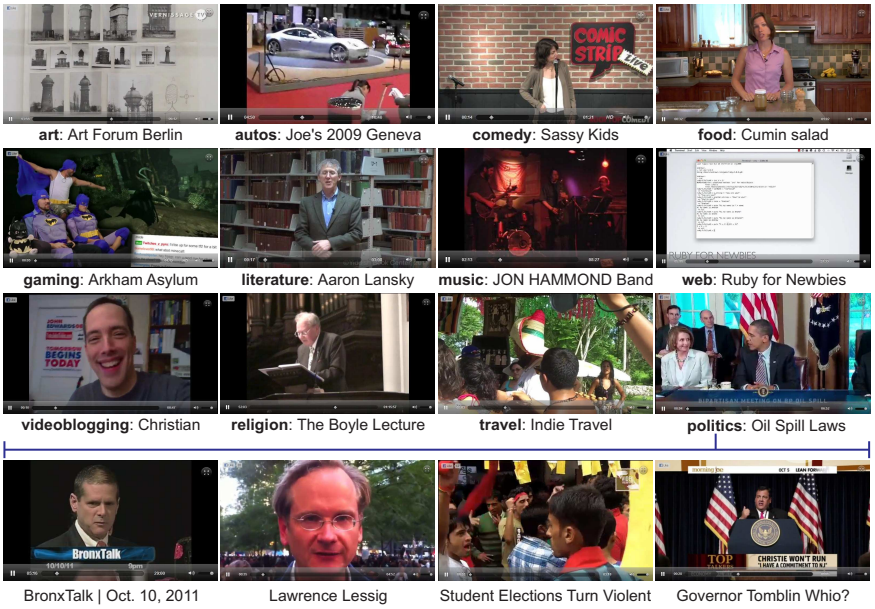



**Fig. 1** Image examples of several video genres. The bottom images exemplify the video diversity of the "politics" category (source blip.tv).

Prior to video processing, we adopt a basic normalization step by converting all sequences to a reference video format. For genre categorization, each movie is represented with a feature vector which corresponds to the previously presented content descriptors. Data fusion is carried out with an early fusion approach [32]. In the following we shall detail each experiment.

### 5.1 A classification perspective

**Experimental setup**. In the first experiment, we address video genre categorization from the perspective of machine learning techniques and therefore we attempt to regroup the data according to genre. For classification we use the

Weka [17] environment which provides a great overview of the existing machine learning techniques. We test methods from simple Bayes, function based, lazy classifiers, rule based, to tree approaches (for each category of methods, we selected the most representative ones). Method parameters were tuned based on preliminary experimentations.

As the choice of training data may distort the accuracy of the results, we use a cross validation approach. The data set is split into train and test sets. We use different values for the percentage split, ranging from 10% to 90%. For a certain amount of training data, in order to shuffle all sequences, classification is repeated for all possible combinations between train and test sets. Additionally, we test different combination of descriptors.

To assess performance we use several measures. At genre level we compute average precision ($P$) and recall ($R$) (averaged over all experimentations for a given percentage split), which account for the number of false classification and misclassifications, respectively, thus:

$$P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}, \quad R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \tag{6}$$

where $\overline{TP}$, $\overline{FP}$ and $\overline{FN}$ represent the *average* number of true positives, false positives and false negatives, respectively. As a global measure, we compute $F_{score}$ and average correct classification ($\overline{CD}$), thus:

$$F_{score} = 2 \cdot \frac{P \cdot R}{P + R}, \quad \overline{CD} = \frac{\overline{N_{GD}}}{N_{total}} \tag{7}$$

where $\overline{N_{GD}}$ is the average number of good classifications and $N_{total}$ is the number of test sequences.

**Discussion on the results**. The most accurate classification is obtained when fusing all audio-visual descriptors together. For reasons of brevity, we limit the presentation to those results. In Figure 2 we present the overall average $F_{score}$ and average correct classification, $\overline{CD}$, for a selection of seven machine learning techniques (the ones providing the most significant results).

The global results are very promising considering the high difficulty of this task. The highest average $F_{score}$ is up to 46.3% while the average correct classification is up to 55% (obtained for 80% training, i.e. from 475 test sequences 261 were correctly labeled). In terms of classification technique, the most accurate proves to be a SVM with linear kernel (see the Black line in Figure 2), followed very closely by Functional Trees (FT, see the Violet line), and further k-NN (with k=3), Random Forest trees, Radial Basis Function Network (RBF), J48 decision tree and finally Bayes Network (see Weka [17]).

The most interesting results are yet obtained at genre level. Due to the high semantic contents, not all genres are to be accurately classified with audio-visual information. We attempt to determine which categories are better suited for this approach. In Figure 3 we present genre average $F_{score}$ obtained with the linear SVM and FT trees.

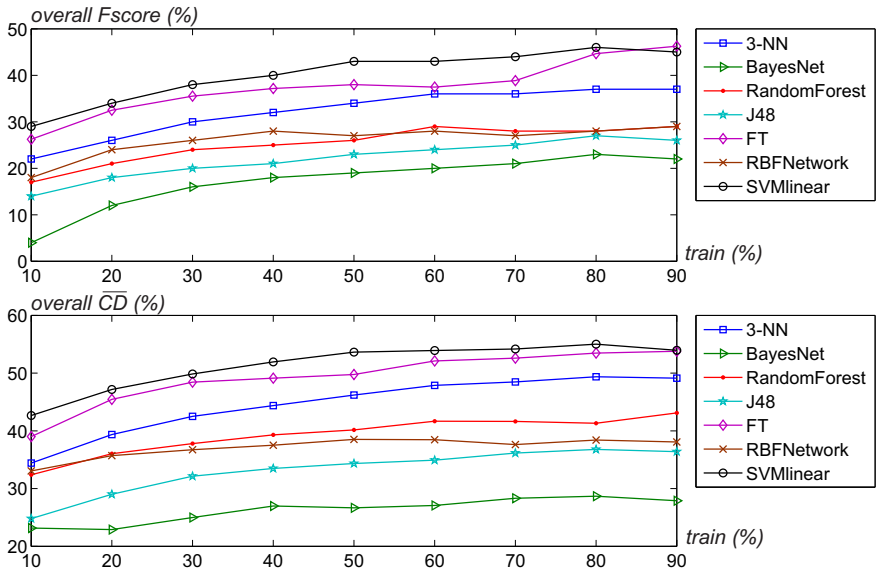



**Fig. 2** Overall average $F_{score}$ (see eq. 7) and overall average correct classification $\overline{CD}$ (see eq. 7) obtained with various machine learning techniques run on all audio-visual descriptors.

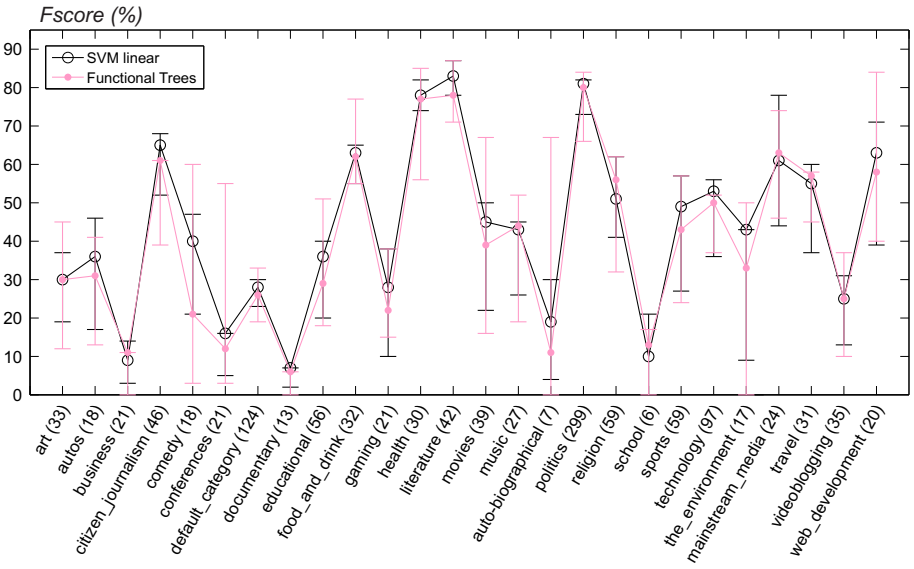



**Fig. 3** Average $F_{score}$ (see eq. 7) for linear SVM and Functional Trees (FT) run on all audio-visual descriptors and for a train-test percentage split of 50% (the number in the brackets represent the number of test sequences for each genre). Vertical lines present for each genre the min-max $F_{score}$ intervals (percentage split ranging from 10% to 90%).

The best performance is obtained for the following genres (we present results for a 50% percent split, as well as the highest value): "literature" ($F_{score} = 83\%$, highest 87%) and "politics" ($F_{score} = 81\%$, highest 84%), followed by "health" ($F_{score} = 78\%$, highest 85%), "citizen journalism" ($F_{score} = 65\%$, highest 68%), "food and drink" ($F_{score} = 62\%$, highest 77%), "web development and sites" ($F_{score} = 63\%$, highest 84%), "mainstream media" ($F_{score} = 63\%$, highest 74%), "travel" ($F_{score} = 57\%$, highest 60%), "technology" ($F_{score} = 53\%$, highest 56%). At the bottom end we have genres like "documentary" ($F_{score} = 7\%$ which is also the highest), "school" ($F_{score} = 10\%$, highest 22%) or "business" ($F_{score} = 9\%$, highest 14%).

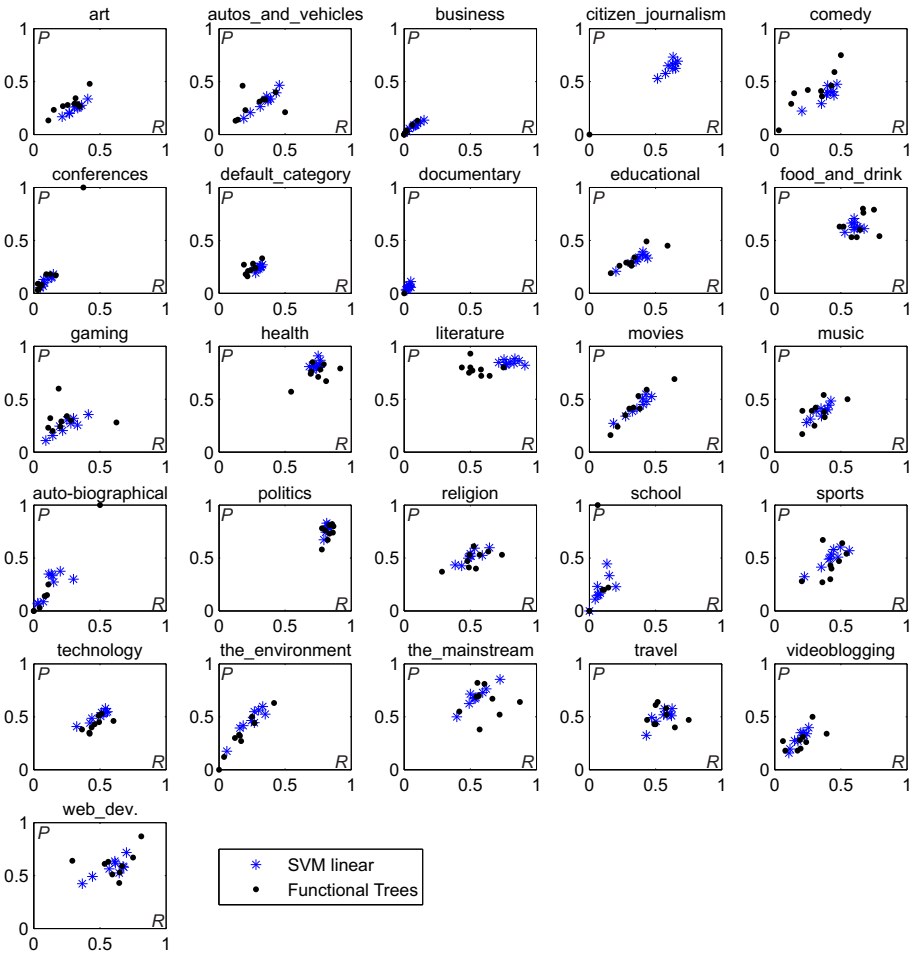



**Fig. 4** Average precision vs. recall (see eq. 6) obtained with SVM (linear kernel) and Functional Trees (FT), run on all audio-visual descriptors and for various amounts of training data (percentage split from 10% to 90%).

Globally, classification performance increases with the amount of training data. However, for some genres, due to the high variety of video materials (see Figure 1), increasing the number of examples may result in overtraining and thus reducing the classification performance. This is visible with Figure 2 where classification performance may decrease with the increase of training (e.g. SVM linear for 90% training).

A clear delimitation between FT and SVM is drown at genre level. Globally, SVM tends to perform better on a reduced training set, while FT tends to outperform for higher amounts of training (e.g. training $> 70\%$, see min-max intervals in Figure 3). This is also visible with genre precision and recall.

In Figure 4 we present genre average precision against recall for various percentage splits (ranging from 10% to 90%). In this case, the highest average precision and thus the lowest number of false classifications is achieved for genres like "literature" ($P = 93\%$ with FT), "health" ($P = 90.9\%$ with FT), "web development and sites" ($P = 87\%$ with FT), "the mainstream media" ($P = 85.3\%$ with SVM), "politics" ($P = 82.9\%$ with SVM), "food and drink" ($P = 79\%$ with FT), "comedy" ($P = 75\%$ with FT), "citizen journalism" ($P = 73\%$ with SVM), "movies and television" ($P = 69\%$ with FT) and "sports" ($P = 67\%$ with FT).

In terms of misclassification, the highest average recall is obtained for "literature" ($R = 91.3\%$ with SVM), "politics" ($R = 86.6\%$ with FT), "the mainstream media" ($R = 87.5\%$ with FT), "web development and sites" ($R = 81.3\%$ with FT), "food and drink" ($R = 79.2\%$ with FT) and "travel" ($R = 75\%$ with FT). Note that most of these values are obtained for the highest amount of training, i.e. 90%.

### 5.2 A retrieval perspective

**Experimental setup**. In this experiment we assess the classification performance of the proposed descriptors from the perspective of an information retrieval system. We present the results obtained at the MediaEval 2011 Benchmarking Initiative for Multimedia Evaluation, the Video Genre Tagging Task [2]. The challenge was to develop such a system able provide the retrieval of all genres. Each participant was provided with a development set consisting of 247 sequences (we eventually extended the data set to up to 648 sequences) and the actual retrieval was performed on 1727 sequences. In this case the training-classification steps are to be performed once. Up to 10 teams competed for this task, each one submitting up to 5 different runs from which 3 were restricted to use only textual descriptors (i.e. provided speech transcripts, user tags, metadata). A detailed overview of the results is presented in [2].

In our case, the retrieval results are provided using a binary ranking, where the maximum relevance of 1 is associated to the genre category to which the document was classified, while other genres have 0 relevance. To assess performance we use the overall Mean Average Precision (MAP) as defined with TRECVid [3] (see also trec_eval scoring tool at `http://trec.nist.gov/`

trec_eval/), thus:

$$MAP = \frac{1}{|Q|} \cdot \sum_{j=1}^{|Q|} \frac{1}{m_j} \cdot \sum_{k=1}^{m_j} P(R_{j,k}) \tag{8}$$

where $Q = \{q_1, ..., q_{|Q|}\}$ represents a set of queries $q_j$ which are represented in the data set with $\{d_1, ..., d_{m_j}\}$ relevant documents, $R_{j_k}$ is the set of ranked retrieval results from the top result until you get to document $d_k$ and $P()$ is the precision (see eq. 6). When a relevant document is not retrieved at all the precision value in the above equation is taken to be 0.

**Discussion on the results**. For the classification we use the approach providing the most accurate results, namely the SVM with a linear kernel. In Table 1 we compare our results against several other approaches using various modalities of the video, form textual information (e.g. speech transcripts, user tags, metadata) to audio-visual[1].

The proposed descriptors achieved an overall MAP up to 12% (see team RAF in Table 1). From the modality point of view, these are the best results obtained using audio-visual information alone. As comparison order, the use of descriptors like cognitive information (face statistics), temporal (average shot duration, distribution of shot lengths) [29], audio (MFCC, zero crossing rate, signal energy), color (histograms, color moments, autocorrelogram - denoted autocorr.) and texture (co-occurrence - denoted co-occ., wavelet texture grid, edge histograms) with SVM led to MAP less than 1% (see team KIT in Table 1); while clustered SURF features and SVM achieved MAP up to 9.4% (see team TUB in Table 1). We obtain better performance even compared to some classic text-based approaches, e.g. the use of Term Frequency-Inverse Document Frequency (TF-IDF) - MAP 9.8% (see team UAB in Table 1) or the use of Bag-of-Words - MAP 5.5% (see team SINAI in Table 1). Compared to visual information, audio descriptors seem to provide better discriminative power to this task.

One should note that these results are however indicative results and a direct comparison may be subjective in some cases, as the results are dependent on both the way the training was carried out as well as on the setting up of the classification approach. The results presented in Table 1 are in general obtained for different training setups (teams were allowed to access other sources of information than the ones proposed during competition). For instance, we use for training 648 sequences compared to team KIT which uses up to 2514 sequences. Most of the text based approaches use in general query expansion techniques (e.g. Wordnet - see http://wordnet.princeton.edu/, Wikipedia - see http://en.wikipedia.org). The interest in showing these results is to provide the reader with a comparison order, as well as to highlight the difficulty of this task.

[1] the following notations were adopted: Terrier IR is an information retrieval system, see http://terrier.org/; Delicious is a social tagging site, see http://del.icio.us/.

**Table 1** Comparative results: MediaEval benchmarking [2] (selective results).

| descriptors | modality | method | decision | MAP | team |
|---|---|---|---|---|---|
| speech transcripts | text | Support Vector Machines | ranked list | 11.79% | LIA |
| speech transcripts, metadata, user tags | text | Bag-of-Words + Terrier IR | ranked list | 11.15% | SINAI |
| speech transcripts | text | Bag-of-Words | ranked list | 5.47% | SINAI |
| speech transcripts, metadata, user tags | text | TF-IDF + cosine dist. | binary | 9.4% | UAB |
| speech transcripts, Delicious tags, metadata | text | BM25F [30] + Kullback - Leibler diverg. | ranked list | 11.11% | UNED |
| metadata | text | Negative multinomial diverg. | ranked list | 39.37% | TUD |
| MFCC, zero cross. rate, signal energy | audio | multiple SVMs | binary | 0.1% | KIT |
| **proposed** | audio | SVM with linear kernel | binary | 10.29% | RAF |
| clustered SURF | visual | Visual-Words + SVM with RBF kernel | binary | 9.43% | TUB |
| hist., moments, autocorr., co-occ., wavelet, edge hist. | visual | multiple SVMs | binary | 0.35% | KIT |
| cognitive (face statistics [29]) | visual | multiple SVMs | binary | 0.1% | KIT |
| structural (shot statistics [29]) | visual | multiple SVMs | binary | 0.3% | KIT |
| **proposed** | visual | SVM with linear kernel | binary | 3.84% | RAF |
| color, texture, aural, cognitive, structural | audio, visual | multiple SVMs | binary | 0.23% | KIT |
| **proposed** | audio, visual | SVM with linear kernel | binary | 12.08% | RAF |
| clustered SURF, metadata | visual, text | Naive Bayes, SVM + serial fusion | binary | 30.33% | TUB |

The competition results prove that the most efficient retrieval approach still remains the inclusion of textual information, as it accounts for a higher semantic level of description than audio-visual. Average MAP obtained including textual descriptors is around 30% (e.g. see team TUB in Table 1). The retrieval performance is boosted up when including information like movie names, movie's ID from *blip.tv* or the username of the video; in this particular case reported MAP is up to 56% (which is also the highest obtained).

In the following experiment we aim to prove that despite the superiority of text descriptors, audio-visual information alone has great potential in approaching this task whether it can benefits from some additional help, i.e. user relevance feedback.

### 5.3 A relevance feedback perspective

**Experimental setup**. In the final experiment we attempt to enhance the retrieval using relevance feedback. For tests, we use the entire data set, thus all the 2375 sequences. The user feedback is to be automatically simulated from the known class membership of each video (i.e. the genre labels). Compared to real user feedback, this has the advantage of providing a fast and extensive simulation framework; which otherwise could not be achieved due to physical constraints (e.g. the availability of an important number of users) and inherent human errors (e.g. indecision, misperception) that a real user could be subject to. Tests are to be conducted for various sizes of the browsing window.

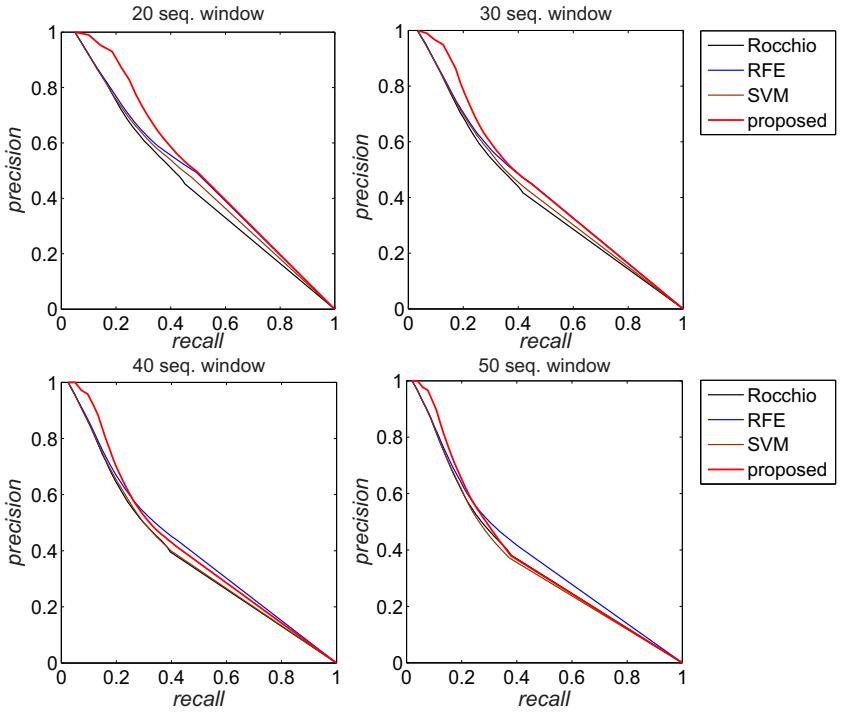



**Fig. 5** Precision - recall curves obtained with relevance feedback applied for different size browsing windows.

**Discussion on the results**. Figure 5 compares the precision - recall curves obtained with the proposed approach, i.e. the hierarchical clustering relevance feedback (HCRF, see Section 4), against several other approaches, namely Rocchio [34], Feature Relevance Estimation (RFE) [36] and Support Vector Machines [37]. One can observe that the proposed HCRF provides an improvement in retrieval compared to the others and in particular for small browsing windows (e.g. 20, 30 video sequences, see the red line in Figure 5). By increas-

ing the window size, all methods tend to converge at some point to similar results.

**Table 2** MAP obtained with Relevance Feedback

| RF method | 20 seq. window | 30 seq. window | 40 seq. window | 50 seq. window |
|---|---|---|---|---|
| Rocchio | 46.8% | 43.84% | 42.05% | 40.73% |
| RFE | 48.45% | 45.27% | 43.67% | 42.12% |
| SVM | 47.73% | 44.44% | 42.17% | 40.26% |
| **proposed** | **51.27%** | **46.79%** | **43.96%** | **41.84%** |

Table 2 summarizes the overall retrieval MAP (see also eq. 8) estimated as the area under the uninterpolated precision-recall curve. For the proposed HCRF, MAP ranges from 41.8% to 51.3% which is an improvement of at least a few percents compared to the other methods. Relevance feedback proves to be an interesting alternative of improving genre retrieval performance, being able to provide results close to the ones obtained with high level textual descriptors.

## 6 Conclusions

In this paper we addressed the issue of web media categorization from the perspective of audio-visual information. We proposed content descriptors which exploit audio, temporal structure and color information and tested their potential to this task. Experimental validation was carried out in the context of the MediaEval 2011 Benchmarking Initiative for Multimedia Evaluation, the Video Genre Tagging Task [2]. It addressed a real scenario, thus the categorization of up to 26 video genres specific to blip.tv media platform (288 hours of video footage). The tests were conducted both from the perspective of a classification system and as an information retrieval approach.

From the classification point of view, not all genres are to be retrieved using audio-visual information. The use of audio-visual information may be highly efficient for detecting some particular genres, like for instance in our case "literature" (we obtain $F_{score} = 87\%$), "politics" ($F_{score} = 84\%$) or "health" ($F_{score} = 85\%$), and less for others, e.g. "school" ($F_{score} = 22\%$) or "business" ($F_{score} = 14\%$). One can imagine a classification system which adapts the parameters to the target categories, like for instance using audio-visual descriptors for genres which are best detected with this information; using text for text-related categories, and so on.

From the retrieval point of view, during MediaEval benckmarking, the proposed descriptors achieved the best results obtained using audio-visual information alone. They provided better retrieval performance than the use of descriptors like cognitive information (face statistics), temporal (average shot duration, distribution of shot lengths), audio (MFCC, zero crossing rate, signal

energy), color (histograms, color moments, autocorrelogram - denoted autocorr.), texture (co-occurrence - denoted co-occ., wavelet texture grid, edge histograms) or even compared to some classic text-based approaches, e.g. the use of Term Frequency-Inverse Document Frequency (TF-IDF). However, the results are still below the ones obtained with text-based descriptors. To compensate this, we designed a relevance feedback approach which allows to boost up the performance close to the one obtained with high semantic textual information (in this particular case, we achieve MAP up to 51%).

The main limitation of this approach, which is also the limitation of all ad-hoc genre categorization approaches, is in the ability of detecting genre related contents. The proposed categorization system is limited in detecting genre related patterns from the global point of view, like episodes from a series being not able to detect a genre related content within a sequence. Therefore, in order to provide good classification performance, each type of video material has to be properly represented with the training set.

Future improvements will mainly consist on approaching sub-genre categorization and considering the constraints of the very large scale approaches (millions of sequences and tens of genre concepts).

## 7 Acknowledgments

This work was supported by the Romanian Sectoral Operational Programme Human Resources Development 2007-2013 through the Financial Agreement `POSDRU/89/1.5/S/62557` and by the Austrian Science Fund (FWF): `L511-N15`. Also, we would like to acknowledge the 2011 Genre Tagging Task of the MediaEval Multimedia Benchmark [2] for providing the test data set.

## References

1. D. Brezeale, D.J. Cook, "Automatic Video Classification: A Survey of the Literature," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(3), pp. 416-430, 2008.
2. M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, G.J.F. Jones (eds.), "Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task", `http://www.multimediaeval.org`, MediaEval 2011 Workshop, Pisa, Italy, 2011.
3. A. F. Smeaton, P. Over, W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements, Multimedia Content Analysis," Theory and Applications, Springer Verlag-Berlin, pp. 151-174, ISBN 978-0-387-76567-9, 2009.
4. G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," Journal of Machine Learning Research, 3, pp. 1289-1305, 2003.
5. M. H. Lee, S. Nepal, U. Srinivasan, "Edge-based Semantic Classification of Sports Video Sequences," IEEE Int. Conf. on Multimedia and Expo, 2, pp. 157-160, 2003.
6. J. Fan, H. Luo, J. Xiao, L. Wu, "Semantic Video Classification and Feature Subset Selection under Context and Concept Uncertainty," ACM/IEEE Conference on Digital Libraries, pp. 192-201, 2004.
7. X. Gibert, H. Li, D. Doermann, "Sports Video Classification using HMMs," Int. Conf. on Multimedia and Expo, 2, pp. II-345-348, 2003.
8. M. Ivanovici, N. Richard, "The Colour Fractal Dimension of Colour Fractal Images", IEEE Transactions on Image Processing, 20(1), pp. 227 - 235, 2010.

9. G. Wei, L. Agnihotri, N. Dimitrova, "TV Program Classification based on Face and Text Processing," IEEE Int. Conf. on Multimedia and Expo, 3, pp. 1345-1348, 2000.
10. X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, Automatic Video Genre Categorization using Hierarchical SVM," IEEE Int. Conf. on Image Processing, pp. 2905-2908, 2006.
11. G. Y. Hong, B. Fong, A. Fong, "An Intelligent Video Categorization Engine," Kybernetes, 34(6), pp. 784-802, 2005.
12. H. Wang, A. Divakaran, A. Vetro, S.-F. Chang, H. Sun, "Survey of Compressed-Domain Features used in Audio-Visual Indexing and Analysis," Journal of Visual Communication and Image Representation, 14(2), pp. 150-183, 2003.
13. Z. Wang, M. Zhao, Y. Song, S. Kumar, B. Li, "YouTubeCat: Learning to Categorize Wild Web Videos", In Proc. of Computer Vision and Pattern Recognition, pp. 879886, 2010.
14. M.J. Roach, J.S.D. Mason, "Video Genre Classification using Dynamics," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1557-1560, Utah, USA, 2001.
15. T. Sikora, "The MPEG-7 Visual Standard for Content Description An Overview," IEEE Trans. on Circuits and Systems for Video Technology, 11(6), pp. 696-702, 2001.
16. B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, P. Lambert, "Content-based Video Description for Automatic Video Genre Categorization", The 18th International Conference on MultiMedia Modeling, 4-6 January, Klagenfurt, Austria, 2012.
17. Weka Data Mining with Open Source Machine Learning Software in Java, University of Waikato, `http://www.cs.waikato.ac.nz/ml/weka/`, 2011.
18. K. Seyerlehner, M. Schedl, T. Pohle, P. Knees, "Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation," 6th Annual Music Information Retrieval Evaluation eXchange (MIREX-10), Utrecht, Netherlands, August 9-13, 2010.
19. X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, "Automatic Video Genre Categorization using Hierarchical SVM," IEEE Int. Conf. on Image Processing, pp. 2905-2908, 2006.
20. P. Kelm, S. Schmiedeke, T. Sikora, "Feature-based video key frame extraction for low quality video sequences", 10th Workshop on Image Analysis for Multimedia Interactive Services, pp.25-28, 6-8 May, London, UK, 2009.
21. W.A.C. Fernando, C.N. Canagarajah, D.R. Bull, "Fade and Dissolve Detection in Uncompressed and Compressed Video Sequence," IEEE Int. Conf. on Image Processing, Kobe, Japan, pp. 299-303, 1999.
22. C.-W. Su, H.-Y.M. Liao, H.-R. Tyan, K.-C. Fan, L.-H. Chen, "A Motion-Tolerant Dissolve Detection Algorithm," IEEE Trans. on Multimedia, 7(6), pp. 1106-1113, 2005.
23. Z. Rasheed, M. Shah, "Movie Genre Classification by Exploiting Audio-Visual Features of Previews," IEEE Int. Conf. on Pattern Recognition, 2, pp. 1086-1089, 2002.
24. U. Srinivasan, S. Pfeiffer, S. Nepal, M. Lee, L. Gu, S. Barrass, "A Survey of Mpeg-1 Audio, Video and Semantic Analysis Techniques," Multimedia Tools and Applications, 27(1), pp. 105-141, 2005.
25. Z. Liu, J. Huang, Y. Wang, "Classification of TV Programs based on Audio Information using Hidden Markov Model," IEEE Workshop on Multimedia Signal Processing, pp. 27-32, 1998.
26. R. Lienhart, "Reliable Transition Detection in Videos: A Survey and Practitioners Guide", Int. Journal of Image and Graphics, 1(3), pp. 469-486, 2001.
27. B. T. Truong, C. Dorai, S. Venkatesh, "Automatic Genre Identification for Content-Based Video Categorization," Int. Conf. on Pattern Recognition, IV, pp. 230-233, 2000.
28. Z. Rasheed, Y. Sheikh, M. Shah, "On the use of Computable Features for Film Classification," IEEE Trans. Circuits and Systems for Video Technology, 15, pp. 5264, 2003.
29. M. Montagnuolo, A. Messina, "Parallel Neural Networks for Multimodal Video Genre Classification", Multimedia Tools and Applications, 41(1), pp. 125-159, 2009.
30. J. Pérez-Iglesias, J. R. Pérez-Agüera, V. Fresno, Y. Z. Feinstein, "Integrating the Probabilistic Models BM25/BM25F into Lucene". CoRR, abs/0911.5046, 2009.
31. B. Ionescu, D. Coquin, P. Lambert, V. Buzuloiu, "A Fuzzy Color-Based Approach for Understanding Animated Movies Content in the Indexing Task," Eurasip Journal on Image and Video Processing, doi:10.1155/2008/849625, 2008.

32. Cees G. M. Snoek, M. Worring, Arnold W. M. Smeulders, "Early versus Late Fusion in Semantic Video Analysis", ACM Int. Conf. on Multimedia, New York, USA, 2005.
33. C.D. Manning, P. Raghavan, H. Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008.
34. N. V. Nguyen, J.-M. Ogier, S. Tabbone, and A. Boucher, "Text Retrieval Relevance Feedback Techniques for Bag-of-Words Model in CBIR", International Conference on Machine Learning and Pattern Recognition, 2009.
35. W. J. Krzanowski, "Principles of Multivariate Analysis: A User's Perspective", Clarendon Press, Oxford, 1993.
36. Yong Rui, T. S. Huang, M. Ortega, M. Mehrotra, S. Beckman, "Relevance feedback: a power tool for interactive content-based image retrieval", IEEE Trans. on Circuits and Video Technology, 8(5), pp. 644-655, 1998.
37. S. Liang, Z. Sun, "Sketch retrieval and relevance feedback with biased SVM classification," Pattern Recognition Letters, 29, pp. 1733-1741, 2008.

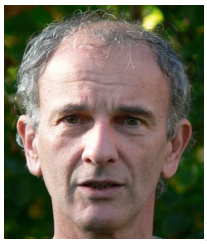
*Author Biographies
Click here to download Author Biographies: Bio.tex

**Bogdan Ionescu** is currently a Lecturer with University "Politehnica" of Bucharest-Romania. He holds a B.S. degree in applied electronics (2002) and an M.S. degree in computing systems (2003), both from University Politehnica of Bucharest. He also holds a Ph.D. degree in image processing (2007) from, both, the University of Savoie and University "Politehnica" of Bucharest. Between 2006 and 2007, he held a temporary Assistant Professor position with Polytech'Savoie, University of Savoie. His scientific interests cover video processing, video retrieval, computer vision, software engineering, and computer science. He is a Member of IEEE, SPIE, ACM, and GDR-ISIS.

**Klaus Seyerlehner** is a postdoctoral researcher at the Department of Computational Perception at the Johannes Kepler University in Linz, Austria. He holds a M.S. (2006) and a Ph.D. (2011) degree in computer science both from the Johannes Kepler University. His main research interests cover the fields of digital music signal processing, pattern recognition, machine learning, statistics and recommender systems.

**Ionuţ Mironică** received the B.S. degree in electrical engineering from Politehnica University of Bucharest in 2009 and the M.S. degree in Databases from the Bucharest Academy of Economic Studies - Cybernetics, Statistics and Informatics Faculty in 2011. He is currently pursuing the Ph.D. degree at the Politehnica University of Bucharest, at LAPI Laboratory. He focused his research on applying interactive pattern recognition and information retrieval techniques to problems of multimedia search, retrieval, and clustering.

**Constantin Vertan** holds an image processing and analysis tenure at the Image Processing and Analysis Laboratory from the Faculty of Electronics, Telecommunications and Information Technology at the "Politehnica" University of Bucharest (UPB). He was an invited professor at INSA de Rouen and University of Poitiers (France). For his contributions in image processing he was awarded with UPB's "In tempore opportuno" award (2002) and with the Romanian National Research Council "In hoc signo vinces" award (2004). His research interests are general image processing and analysis, CBIR, fuzzy and medical image processing applications. He is a member of SPIE, senior member of IEEE and secretary of the Romanian IEEE Signal Processing Chapter.



**Patrick Lambert** received the enginer degree in electrical engineering in 1978, and the PhD degree in signal processing in 1983, both from the National Polytechnic Institute of Grenoble, France. He is currently a Full Professor at the School of Engineering of University of Savoie, Annecy, France and a member of the Informatics, Systems, Information and Knowledge Processing Laboratory (LISTIC), Annecy, France. His research interests are in the field of image and video analysis, and actually dedicated to non linear color filtering and video semantic indexing.